An Introduction to Spike-and-Slab priors

A useful tool for high-dimensional regression

Michael Komodromos

October 18, 2021

Outline

- Motivation - why should we care?
- Spike-and-Slab priors - what they are?
- A case study - how do we use them?

Summary

Spike-and-slab priors are a useful tool for high-dimensional regression.
Demonstrating better performance than traditional methods (e.g. LASSO)

Motivation

We have a response of interest: continuous number, class type, survival time.

We want to know how are features are associated with the response

But ...

The features we're analysing are high-dimensional ($p \gg n$)

So it's hard to decide which ones are associated with our response

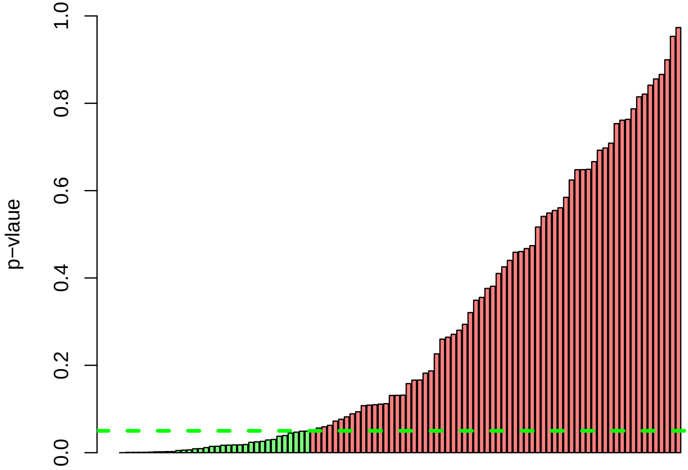
What's being used?

Univariate selection

- We fit a regression model for each feature
- Compute a p-value
- After setting a threshold and include / exclude variables

What's being used?

Univariate selection



What's being used?

LASSO

Figure: Rob Tibshirani

What's being used?

LASSO

- A popular method for penalised regression
- Commonly used when we have high dimensional data
- Returns point estimates
- Can't really do uncertainty quantification without further methods

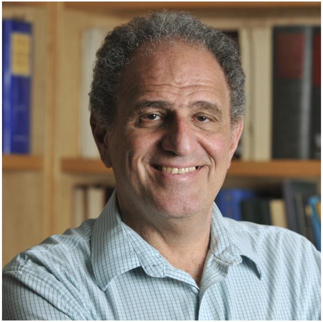


Figure: Rob Tibshirani

What's being used?

LASSO, some extra details

Returns $\widehat{\beta}$

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}} \;\; \ell(\beta; y, X) - \lambda|\beta|$$

where y is our response, X our design matrix and β our model coefficients.

What we would like?

Ideally we would like a combination of univariate selection and the LASSO, i.e. we want to know:

- the probability a feature is associated with our response
- the effect size, $\widehat{\beta}$

And as an added bonus, uncertainty quantification

The solution

Spike and slab priors

A brief history

- Appeared in the 1980s, predating the LASSO! See [MB88; GM93] for theory and examples
- But - these are Bayesian methods, i.e. computationally expensive

Figure: High end computer from 1980

Spike-and-Slab priors

Each coefficient β_j within our model has a corresponding latent variable z_j

z_j indicates whether the coefficient takes a value of 0 or not.

Formally,

$$w_j \overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \tag{1}$$

$$z_j | w_j \overset{\text{ind}}{\sim} \text{Bernoulli}(w_j) \tag{2}$$

$$\beta_j | z_j \overset{\text{iid}}{\sim} z_j \text{Laplace}(\lambda) + (1 - z_j)\text{Dirac}_0 \tag{3}$$

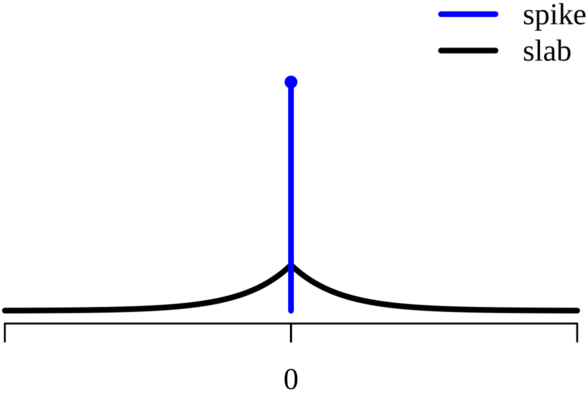
Spike-and-Slab priors

And we update our prior inclusion probabilities w_j and coefficient value β_j using Bayes theorem

$$\Pi(\beta, z, w|\mathcal{D}) \propto L(\mathcal{D}; \beta, z, w)\Pi(\beta, z, w) \tag{4}$$

where L is our likelihood, Π our prior and $\Pi(\cdot|\mathcal{D})$ our posterior.

Spike-and-Slab priors



The good

- Addresses all our problems
- Not as sensitive to prior parameters
- Lower false discovery rate than other methods
- Describes all possible models

$$
\begin{array}{ccccccc}
z_1 & z_2 & z_3 & \cdots & z_{p-1} & z_p & \\
\boxed{1} & \boxed{0} & \boxed{0} & \cdots & \boxed{0} & \boxed{0} & \mathcal{M}_1 \\
\boxed{0} & \boxed{1} & \boxed{0} & \cdots & \boxed{0} & \boxed{0} & \mathcal{M}_2 \\
\vdots & & & & & \vdots & \\
\boxed{1} & \boxed{1} & \boxed{1} & \cdots & \boxed{1} & \boxed{1} & \mathcal{M}_{2^p}
\end{array}
$$

The bad

- Difficult to implement - can't be done in `stan`
- Computationally expensive

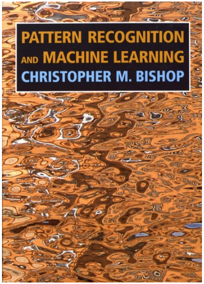
The ugly, a solution to the bad

Rather than computing a perfect solution, we compute an approximation.

This comes at a cost, our approximation is overconfident

Some more details

- We're using variational inference to approximate the posterior
- See [Bis06; BKM17]





A case study

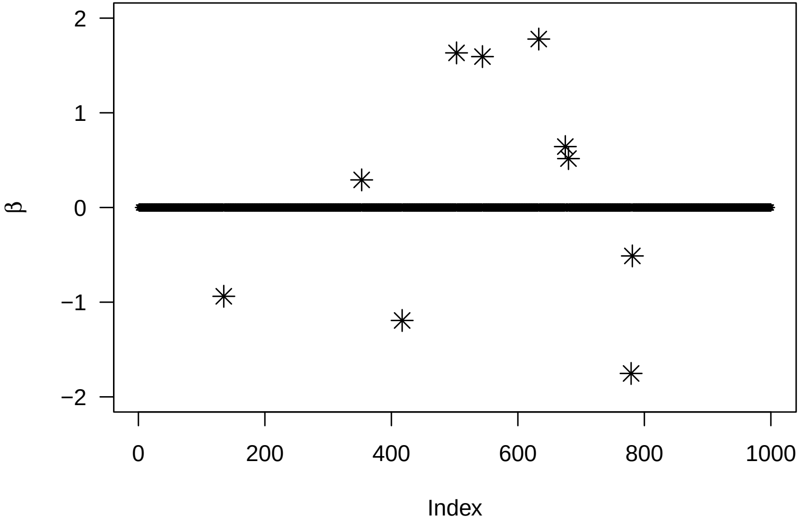
Recently, I've been extending Spike + Slab priors to the Cox model

We're going to be looking at survival data where we have

- $n = 200$ (with 40% of times censored)
- $p = 1,000$

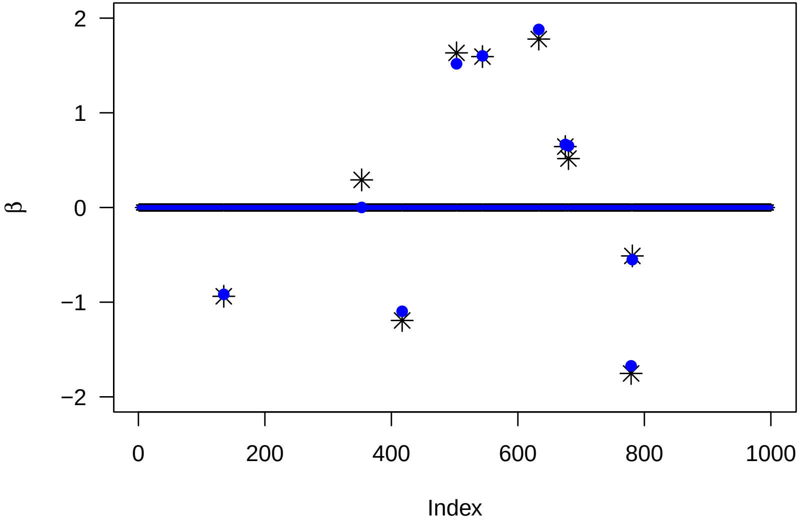
A case study

The true β is



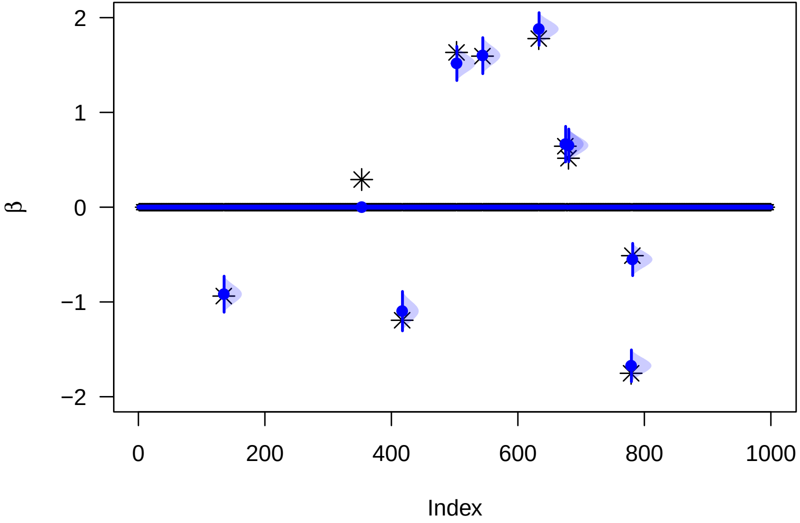
A case study

Our aim: estimate β . We ran the method and ...



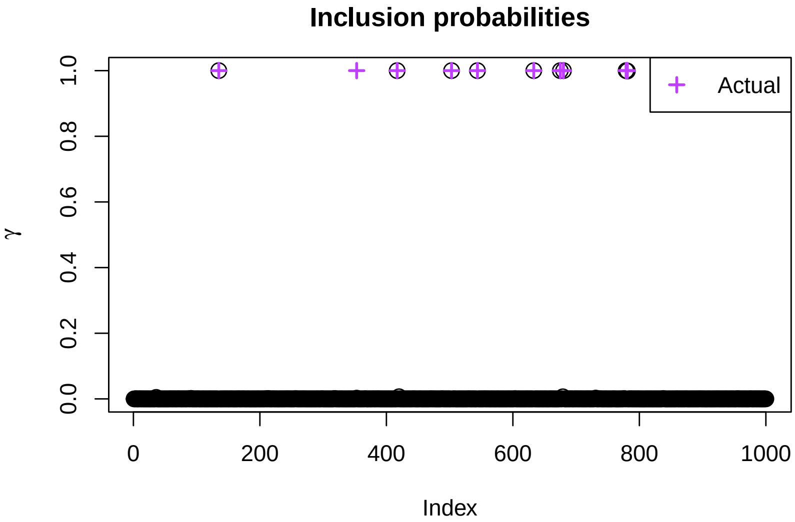
A case study

Not too bad,but there's more



A case study

And the inclusion probabilities,



A case study

Some downsides

- the method is overconfident
- the credible intervals are much tighter than those from MCMC
- small effect sizes aren't picked up
- high-correlation / censoring reduces performance

For next time

More details about

- how the method works
- a comparison to the LASSO + other methods
- a case study on a real dataset

Show me the code

`R` package available at

Variational Bayes for survival

`https://github.com/mkomod/survival.svb`

MCMC sampler

`https://github.com/mkomod/survival.ss`

Reference

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. Chap. 10.

[BKM17] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518 (2017), pp. 859–877.

[GM93] Edward I. George and Robert E. McCulloch. "Variable Selection via Gibbs Sampling". In: *Journal of the American Statistical Association* 88.423 (1993), pp. 881–889.

[MB88] T. J. Mitchell and J. J. Beauchamp. "Bayesian variable selection in linear regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

Questions, Comments?